Toward an Expanded Genome: Structural and Computational Characterization of an Artificially Expanded Genetic Information System

Nigel G. J. Richards[1,2] and Millie M. Georgiadis[3,4]*

[1]School of Chemistry, Cardiff University, Cardiff, CF10 3AT, United Kingdom, [2]Foundation for Applied Molecular Evolution , 13709 Progress Boulevard, Alachua, Florida, 32615, [3]Department of Biochemistry & Molecular Biology, Indiana University School of Medicine, Indianapolis, IN 46202, and [4]Department of Chemistry and Chemical Biology, Indiana University Purdue University at Indianapolis,

* To whom correspondence should be addressed. Tel: 1(317)278-8486, Fax: 1(317)274-4686, email: mgeorgia@iu.edu

This is the author's manuscript of the article published in final edited form as:

Richards, N. G. J., & Georgiadis, M. M. (2017). Toward an Expanded Genome: Structural and Computational Characterization of an Artificially Expanded Genetic Information System. Accounts of Chemical Research, 50(6), 1375–1382. https://doi.org/10.1021/acs.accounts.6b00655

CONSPECTUS

Although the fundamental properties of DNA as first proposed by Watson and Crick in 1953 provided a basic understanding of how duplex DNA was organized and might be replicated, it was not until the first crystal structures of DNA (Z-DNA in 1979, B-DNA in 1980, and A-DNA in 1982) that the true complexity of the molecule began to be appreciated. Many crystal structures of oligonucleotides have since shed light on the helical forms that "Watson-Crick" DNA can adopt, their associated groove widths, and the properties of the nucleobase pairs and their interactions in all three helical forms. Additional understanding of the properties of Watson-Crick DNA has been provided by computational studies employing a variety of theoretical methods. Together with these studies devoted to understanding Watson-Crick DNA, recent efforts to expand the genetic alphabet have founded a new field in synthetic biology. One of these efforts, the artificially expanded genetic information system (**AEGIS**) developed by Steven Benner and coworkers, takes advantage of orthogonal hydrogen bonding to produce DNA comprised of six nucleobase pairs, of which the most extensively studied is referred to as **P:Z** with **P** being 2-amino-imidazo[1,2-a]-1,3,5-triazin-4(8H)-one) and **Z,** 6-amino-5-nitro-2(1H)-pyridone. **P:Z** forms three edge-on hydrogen bonds that differ from standard Watson-Crick pairs in the arrangement of acceptors and donor groups; **P** presents acceptor, acceptor, donor and **Z**, donor, donor, acceptor. **Z** is unique among the **AEGIS** nucleobases in having a nitro group present in the major groove. **PZ**-containing DNA has been exploited in a number of clinical applications, and is being used to develop receptors and catalysts. Ultimately, the grand challenge will be to create a semi-synthetic organism with an expanded genome. And, just as our understanding of the properties of natural DNA have benefited from structural and computational characterization, so too will our understanding of artificial DNA.

This Account focuses on the structural and biophysical properties of **AEGIS** DNA containing **P:Z** pairs. We begin with the fundamental properties of **P:Z** nucleobase pairs, including their electrostatic potential and hydrogen-bonding energies, as elucidated by quantum mechanical calculations. We then examine the impact of including multiple consecutive **P:Z** pairs into duplex DNA providing an opportunity to investigate stacking interactions between **P:Z** pairs. The self-complementary 5'-CTTAT**PP**TA**ZZ**ATAAG was crystallized in B-form using the host-guest system along with analogous natural sequences including G's or A's. Use of the host-guest system to characterize B-DNA obviates a number of limitations on the structural characterization of sequences of interest; these include the ability to crystallize the desired sequences and to distinguish structural effects imparted by the lattice constraints from those inherent in the sequence itself. On the other hand, 3/6ZP, 5'-CTTAT**PPPZZZ**ATAAG, was crystallized in A-form in a DNA-only lattice allowing a comparative analysis of **P:Z** pairs in two of the biologically relevant helical forms, A- and B-DNA. Computational studies on the 3/6ZP sequence starting in A-form provide additional evidence for a more energetically favourable stacking interaction, which we term the "slide" conformer, observed in the A-form crystal structure; this unusual stacking interaction plays a major role in altering the conformational dynamics observed for the **PZ**-containing duplex, as compared to a GC-containing "control" duplex, in long timescale molecular dynamics simulations. This combined use of structural and computational strategies paves the way for obtaining a detailed description of artificial DNA, both in how it differs from Watson-Crick DNA and in the rational discovery of proteins, such as endonucleases, transcription factors and polymerases, which can specifically manipulate DNA containing **AEGIS** nucleobase pairs.

ARTIFICIAL DNA

One of the most important outcomes of modern synthetic biology is the recognition that the biopolymers resulting from four billion years of biological evolution are not the only molecules that might support genetics, inheritance, evolution, and catalysis.[1] As a consequence of efforts to create artificial DNA and thereby an expanded genome capable of Darwinian evolution, we now have a much greater understanding of the fundamental properties of natural DNA. Although the phosphodiester backbone has been shown to be essential for DNA structure and function, the nucleobases can and have been broadly altered resulting in pairs that rely on orthogonal hydrogen bonding patterns or those that dispense with hydrogen bonding and rely on steric and hydrophobic complementarity.[2-6] The most extensive and widely used of these is an artificially expanded genetic information system (**AEGIS)** that employs orthogonal hydrogen bonding patterns between big purine-like and small pyrimidine-like nucleobases maintaining a duplex DNA structure.[2,7] Thus, the number of nucleobase pairs can be increased from two to six by merely rearranging the pattern of hydrogen bond donor and acceptor groups.

Practical efforts to implement this idea using chemical synthesis have thus far led to several “generations” of novel heterocycles that can be employed in automated DNA synthesis, thereby yielding artificially expanded alphabets. DNA molecules containing the **P:Z** nucleobase pair (Fig. 1) have proven of especial interest given that several polymerases will replicate them in nested PCR reactions[8], and Taq DNA polymerase variants have been obtained that will efficiently and faithfully replicate this base pair.[9] A unique feature of **Z** is the nitro group in the major groove providing additional functionality as compared to a natural nucleobase. **P:Z** and possible mismatches have been extensively studied through UV absorbance melting measurements and determination of the energetics of binding of DNA strands providing evidence

that **P:Z** pairs contribute more to duplex stability than any mismatches involving either nucleobase with natural nucleobases and that **P:Z** is more stable than the most similar natural pair G:C.[10] In addition, protein engineering using a variety of strategies has created polymerases that copy and amplify oligonucleotides containing another **AEGIS** pair, **X** and **K** (Fig. 1).[11]

On the other hand, exploiting altered patterns of hydrogen bonding to obtain novel nucleobase pairs that are "orthogonal" to **A**:**T** and **G**:**C** has proven surprisingly problematic.[2,12] For example, many heterocycles have populated tautomeric forms with altered hydrogen bonding patterns; these can base pair with standard nucleobases in either duplex DNA or within the active sites of polymerases, thereby giving rise to unanticipated mutations or the loss of the **AEGIS** nucleobases during replication.[13] Even if these "design" problems were to be easily resolved, little is known about how the incorporation of these non-natural nucleobases into DNA affects the conformational preferences and dynamical properties of this complex molecule, which is fundamental to the interaction of "standard" DNA with proteins, such as polymerases and transcription factors. Indeed the first studies aimed at understanding how **P**:**Z** nucleobase pairs, which have altered electrostatic properties (dipole moments, charge distribution), might perturb DNA structure and dynamics have only recently appeared.[14,15] Moreover, molecular insights into how proteins might recognize **AEGIS**-DNA molecules have not yet been reported, and we note that predictive computational assessments of the interaction free energies between **AEGIS**-based DNA and proteins will considerably aid efforts to create reagents for using **AEGIS** molecular components in bacterial cells, with all of the associated implications for synthetic biology.[16] While many challenges remain, non-natural nucleobase pairs have now been successfully replicated in a bacterial system, paving the way for the development of semisynthetic organisms.[16,17]

QUANTUM MECHANICAL CHARACTERIZATION OF THE P:Z PAIR

As a prelude to examining the global impact of multiple consecutive **P:Z** nucleobase pairs on the DNA duplex, we examined the electrostatic potential and hydrogen bonding properties of the two non-natural nucleobases. Considerable precedent exists in using quantum mechanical (QM) calculations for modeling the electronic properties of individual Watson-Crick nucleobases and their associated base pairs, especially in combination with explicit and continuum solvation models. For example, such computational studies have established the energies of A:T and G:C hydrogen bonding, both *in vacuo* and in aqueous solution,[18,19] and rationalized the energetic preference for unsymmetrical hydrogen bond arrangements,[20] as in A:T and G:C, rather than the symmetrical pattern that is present in X:K nucleobase pairs. To date, QM calculations have only been reported for the **P:Z** nucleobase pair.[14,15,21] Although high-level *ab initio* calculations suggest that the free energy of hydrogen bonding in the **P:Z** nucleobase pair is less favorable (1.4 kcal/mol) than for G:C in the gas-phase,[21] calculations that include the effects of solvation indicate that the presence of the Z-nitro group decreases the enthalpy of the hydrogen bonds in **P:Z** relative to those in G:C by 0.6 kcal/mol.[15,21] This finding is supported by recent biophysical studies in which the **P:Z** hydrogen bonding interactions were found to be stronger than natural or mispaired interactions involving either **P** or **Z**.[10]

Little systematic computational work has been performed to examine how placing **AEGIS** nucleobases within the DNA duplex modifies their electronic distributions, but simple gas-phase QM calculations on isolated **P:Z** and G:C nucleobase pairs clearly show significant differences in dipole moments and electrostatic potential within both grooves (Fig. 2),[21] which can potentially be exploited in re-engineering the specificity of transcription factors and restriction endonucleases.[22]

STRUCTURAL CHARACTERIZATION OF P:Z PAIRS IN B-DNA

General Considerations

Crystallization of a polyanionic molecule like DNA poses a significant challenge due to the limited number of sites available for intermolecular contacts that are required to form a three-dimensional lattice.[23] Thus, the ability to crystallize DNA alone and specifically in B-form has in the past been limited to specific sequences and lengths of oligonucleotides, with the first example of a B-form DNA structure being the Drew-Dickerson dodecamer.[24] The majority of structural studies on B-form DNA have been done on oligonucleotides that are 12 base pairs or shorter in length. The problem of crystallizing DNA is compounded by inclusion of non-natural nucleobases, which represent unchartered territory for structural analyses. Some of these problems can be circumvented through the use of a host-guest system developed by Georgiadis and coworkers in which the N-terminal fragment of Moloney murine leukemia virus reverse transcriptase (MMLV RT) serves as the host and a self-complementary 16-mer oligonucleotide as the guest.[25] In the complex, one DNA duplex is bound to two protein molecules; protein-DNA interactions involve the terminal three nucleobase pairs, with R116 bound in the minor groove and other interactions involving backbone atoms (Fig. 3).To date, this system has been used to determine 24 natural and 2 artificial DNA oligonucleotides structures including some with ligands bound to the DNA (Table 1). We recognized that this type of complex could in fact serve as a host-guest system; the binding site for DNA within the protein is general and could potentially accommodate any 16-mer DNA sequence. The unique repeating unit or the asymmetric unit of the crystal includes only one protein molecule and half of the DNA molecule (Fig. 3). Thus, the system is best suited to self-complementary 16-mer DNA oligonucleotides but

has been successfully used to analyze the structure of sequences that are not self-complementary.[26]

The utility of the host-guest system extends to the crystallization and analysis of DNA sequences of interest[25,27] as well as DNA-ligand complexes[28]. The host-guest system has three major advantages over DNA-only systems; it allow us to (i) crystallize any 16-mer DNA oligonucleotide sequence that adopts B-form, (ii) phase the structure of the complex using the host as a search model in molecular replacement calculations providing unbiased electron density for the DNA, and (iii) analyze DNA structures, determined typically at 1.7-1.8 Å, that have been obtained in the same lattice and are therefore subject to the same constraints allowing DNA sequence-specific features to emerge. Crystals of desired DNA sequences are grown in a low salt, PEG 4000-containing precipitant and can be obtained rapidly through microseeeding with seeds created from crystals of a standard host-guest complex.[27,29] The major limitation of the system is that 16-mer oligonucleotide duplexes must be used to obtain crystals and that 16-mer oligonucleotide must exist predominantly as B-form DNA.

Table 1: **DNA sequences analyzed using the host-guest system**

PDB ID	Type	DNA Sequence
4XN0	AEGIS 3/6 ZP	5'-CTTAT**PPPZZZ**ATAAG[30]
4XO0	AEGIS 2P	5'-CTTAT**PP**TA**ZZ**ATAAG[30]
4XPC		5'-CTTATAAATTTATAAG[30]
4XPE		5'-CTTATGGGCCCATAAG[30]
4M94	Spore product*	5’-ATCCGttATAACGGAT[31]
4M95		5’-ATCCGTTATAACGGAT[31]
2R2R		5’-ATTAGTTATAACTAAT[32]

2R2S	Full bleo B2	5’-ATTAGTTTAACTAAT[32]
2R2T		5’-ATTTAGTTAACTAAAT[32]
2R2U	Partial bleo B2	5’-ATTTAGTTAACTAAAT[32]
3FSI	MG lig 4,4’ bIP	5’-CTTAATTCGAATTAAG[33]
2FJV	MG lig RT29	5’-CTTAATTCGAATTAAG[33]
2FJW		5’-CTTAATTCGAATTAAG[28]
2FJX	MG lig RT29	5’-CTTGAATGCATTCAAG[33]
1ZTT	MG lig netropsin	5’-CTTAATTCGAATTAAG[34]
1ZTW		5’-CTTAATTCGAATTAAG[34]
1N4L	HIV PPT	5’-CTTTT**T**AAAA**G**AAAAG[26]
2FVP	LTR	5’-TTTCATTGCAATGAAA[27]
2FVQ	LTR	5’-CTTTCATTAATGAAAG[27]
2FVR	LTR	5’-TCTTTCATATGAAAGA[27]
2FVS	LTR	5’-CACAATGATCATTGTG[27]

*tt refers to the spore product thymine dimer; MG lig, minor groove binding ligand; bleo B2, bleomycin B2; bIP, 4,4'-Bis(imidazolinylamino)diphenylamine; PPT, polypurine tract; LTR, long terminal repeat. The HIV PPT sequence shown is for one of the two strands that made up this duplex.

Structural properties of 5'-CTTATPPTAZZATAAG

The host-guest system is ideally suited to the rapid analysis of artificial DNA. Thus, the most recent application of the host-guest system has been the crystallization and analysis of artificial DNA including **P:Z** nucleobase pairs (Fig. 3).[30] Although we could easily have generated structures including a single **P:Z** nucleobase pair within a natural DNA environment, we were much more interested in the effects of including consecutive **P:Z** nucleobase pairs on the stacking interactions and overall structural properties of the DNA to determine whether they

had special properties. A-tracts are known to confer rigidity to the structure of DNA,[26] while having multiple G's spaced appropriately enables the formation of quadruplex structures.[35] Thus, we pursued the structural characterization of self-complementary 16-mer sequences including either two consecutive **P:Z** pairs (2P, 5'-CTTAT**PP**TA**ZZ**ATAAG) or six consecutive **P:Z** pairs (3/6ZP, 5'-CTTAT**PPPZZZ**ATAAG).[30] The 2P sequence crystallized in the host-guest system in B-form. The 3/6ZP oligonucleotide did not suggesting that another useful property of the host-guest system is to discriminate between sequences that stably adopt B-form and those that exist in solution as a mixture of different helical forms. This conclusion is supported by the molecular dynamics studies discussed below. We also found this to be true for **X:K** nucleobase pairs. A 16-mer including two consecutive **X:K** nucleobase pairs crystallized readily in the host-guest system, while one including six consecutive **X:K** pairs did not. Oligonucleotides including 5'-GGGCCC or 5'-AAATTT sequences replacing the ZP-containing regions readily crystallized in the host-guest system (Table 1). Analysis of all of these oligonucleotides in low salt conditions by CD suggested that the **P:Z** sequences adopt structures that resemble those of the **G:C** control oligonucleotide.[30] 2P to date is the only structure of **P:Z** pairs in B-form DNA. A crystal structure of a 9-mer oligonucleotide, 5'-G 5-MeSedUGT-Z-ACAC-3' and complementary 5'- G 5-MeSedUGT-P-ACAC-3' including Se modified nucleotides, crystallized with 4 molecules in the asymmetric unit, which are either partially or fully A-form.[36] In this case, there are no reference structures including G:C or A:T crystallized in the same lattice available for comparison; thus, it is difficult to draw any specific conclusions regarding the impact of including **P:Z** in this sequence.

Individual **P:Z** base pairs at positions 6 and 7 in the oligonucleotide sequence (see Fig. 3 for numbering scheme), more closely resemble G:C than A:T pairs located in the same position

in the control host-guest complexes with similar shear, stretch, stagger, and propeller values as assessed by 3DNA (Fig. 4).[37] The **P:Z** pair differs from both G:C and A:T pairs at position 6 in that its buckle angle is -11.9° as compared to -5.6° for G:C and -1.40° for A:T at the same position (Fig. 4) and presents a unique pattern of electronegative atoms in the major groove (Fig. 5). Of particular interest is the zwitterionic nitro group at position 5 of the pyrimidine-like heterocyclic ring (Fig. 1), which provides additional functionality in the major groove. Within the minor groove, **P:Z** presents hydrogen bond acceptors O2 from the pyrimidine-like Z and N3 from the purine-like P as found in all natural base pairs (Figs. 1, 5). The major groove width associated with the two **P:Z** pairs (18.7 Å) is on average 0.7 Å wider than for the G:C pairs and 0.4 Å narrower than that observed for A:T pairs at the same positions as calculated in 3DNA.[37] The minor groove width for the **P:Z** pairs (12.5 Å) is very similar on average to that observed for G:C (12.4 Å) and much wider than that observed for A:T (9.7 Å).

STRUCTURAL ANALYSIS OF P:Z PAIRS IN A-DNA

The 3/6ZP sequence was crystallized in a DNA-only lattice in A-form under high salt conditions (Fig. 6).[30] This is to date the longest oligonucleotide crystallized independently in A-form with the next longest being a G:C rich 14-mer oligonucleotide. The 3/6ZP structure was solved by experimental phasing methods as there was no available model for molecular replacement. The following oligonucleotide including two 5'-BrU (**B**), one per strand, was used to determine the structure, 5'-CT**B**ATPPPZZZATAAG. The crystal was grown in 10 mM magnesium acetate, 50 mM MES pH 5.6, and 1.7 M ammonium sulfate and was cryo-cooled in a solution including the reservoir with 20% glycerol added. The structure was determined by experimental bromine single wavelength anomalous diffraction phasing methods and served as the starting model for molecular dynamics studies of DNA including multiple consecutive **P:Z** pairs.[30]

The properties of the 3/6ZP DNA structure containing multiple consecutive **P:Z** nucleobase pairs were analyzed using 3DNA.[37] The first significant finding was that consecutive non-natural **P:Z** nucleobase pairs can be accommodated in canonical A- and B-helical forms of DNA. As in B-form, the **P:Z** pairs exhibit on average a wider major groove than the G:C structure used for comparison with an average major groove width of 18.9 Å versus 18.0 Å on average. In both structures, **P:Z** formed three hydrogen bonding interactions with distances between heteroatoms typical of those in natural nucleobase pairs. The **P:Z** pairs in B-form DNA were sheared as were equivalent G:C pairs (Fig. 4), while those in the A-form structure were not sheared, similar to G:C pairs in A-form DNA (Fig. 7). Two different types of stacking interactions for PP/ZZ dinucleotide steps were identified in the 2P vs. the 3/6 ZP structures. In the B-form 2P structure, consecutive **Z'**s are stacked in much the same manner as natural nucleobases in a shifted arrangement, while in the A-form 3/6ZP structure, the nitro group of **Z** stacks above the ring of the adjacent **Z**. This stacking arrangement involves a sliding motion comparable to that seen for **G:C** steps in A-form DNA (Fig. 8). These unique structural features suggested to us that it would be of interest to perform computational studies and compare the behavior of the 3/6ZP oligonucleotide to that of a related GC sequence.

QUANTUM MECHANICAL CHARACTERIZATION OF P:Z STACKING

Of course, stacking interactions between adjacent nucleobase pairs contribute to helix stability and ample precedent exists for using QM calculations to obtain estimates of stacking energies for adjacent Watson-Crick nucleobase pairs.[38] We, and others,[14,15] have therefore used these methods to demonstrate that stacking **P:Z** nucleobase pairs is energetically preferred to stacking G:C nucleobase pairs by approximately 2.0 kcal/mol,[21] primarily because of favorable electrostatic interactions between the electron-deficient **Z** ring and the π-electrons of the adjacent

P nucleobase (Fig. 2).[14,21] Perhaps more importantly for duplex stability and conformational properties for DNA containing multiple consecutive **P:Z** nucleobase pairs, high-level QM calculations suggest that the PP/ZZ dinucleotide (i.e. two consecutive **P:Z** pairs) can adopt two different structural forms (Fig. 8), which we have termed "slide" and "shift" conformers because one of the two nucleobase pairs is displaced along the axis corresponding to slide or shift, respectively.[21] Comparison with X-ray crystal structures shows that the slide conformer is similar to what is observed for the stacking of PP/ZZ dinucleotides in A-form DNA while the shift conformer resembles that observed in B-form DNA. The calculated energy difference between the two structures (1.5 kcal/mol) suggests that the "slide" conformer, which features "staggered" stacking of the nitro groups (Fig. 8), is more stable but both conformers are accessible at room temperature and above. Similar QM studies for the GG/CC dinucleotide suggest a greater energetic preference for the slide conformer, which is consistent with the experimental finding that duplexes containing only G:C nucleobase pairs prefer to exist in the A-form structure.[39] Of course, access to viable cells containing an expanded genetic alphabet requires not only that DNA duplexes containing multiple **P:Z** nucleobase pairs adopt canonical helical structures but also that **PZ**-containing sequences can form specific interactions with proteins, such as transcription factors and repressors, in the major groove.[40]As a consequence, redesigning DNA-binding proteins to recognize AEGIS nucleobases requires an understanding of the electrostatic properties of the **P:Z** nucleobase pair within the duplex.

MODELING THE STRUCTURE AND DYNAMICS OF AEGIS DNA

Elucidating the intrinsic properties of individual **AEGIS** nucleobases using QM calculations is a necessary element of developing simplified models, such as force fields, which relate changes in molecular geometry to potential energy. Given the intrinsic biological

importance of understanding how DNA sequence might impact conformational preferences,[41] flexibility and the motional properties of the double helix, high-quality force field parameters now exist to model all of the standard Watson-Crick nucleobases in their lowest energy tautomeric form.[42-44] In addition, there have been systematic studies on the interaction of DNA with counter-ions,[45] such as Na^+, and modeling the electrostatic properties of this highly charged biopolymer using classical potential energy functions.[46] This very large body of work, together with long timescale molecular dynamics (MD) simulations, has resulted in a detailed understanding of duplex DNA structure in water, and the dynamical motions that mediate conformational transitions.[47]

In contrast, there have been almost no reports of MD simulations of duplex DNA built from **AEGIS** nucleobases,[15,21] perhaps because of a lack of well-tested force field parameters for these novel nucleobases. In order to understand the molecular basis for the unique structural features seen in the crystal structures of **PZ**-containing DNA duplexes, we therefore developed parameters for these two **AEGIS** nucleobases assuming that they exist only in their lowest energy tautomeric form (Fig. 1). With these in hand, we were then able to examine the motions and conformational properties of the 3/6ZP oligonucleotide in water over a 50 μsec timescale. This calculation showed that the presence of the six consecutive **P:Z** nucleobase pairs gave rise to a duplex that exhibited structural features associated with both A- and B-form DNA. For example, the 3/6ZP oligonucleotide featured a wider major and narrower minor groove (average values of 27 Å and 13 Å, respectively) over the course of the MD simulation than in the A-form helix observed in the X-ray crystal structure (values of 19 Å and 16.5 Å, respectively).[30] The differences in groove width for the 3/6ZP oligonucleotide in water are, of course, one of the defining features of the B-form double helix. On the other hand, the range of values of structural

measures associated with PP/ZZ dinucleotide steps are those expected for A-form duplex DNA.[21] This conformational behavior is likely associated with adjacent **P:Z** nucleobase pairs interconverting between the "slide" and "shift" conformers identified by QM calculation. In addition, when compared with the dynamical behavior of a "control" DNA duplex in which **P:Z** nucleobase pairs in the 3/6 ZP oligonucleotide are all replaced by G:C, the PZ-containing DNA duplex in water accesses a larger number of conformations over the course of the MD simulations (Fig. 9). These observations contrast sharply with conclusions from an MD simulation of a 15-bp DNA duplex containing only a single **P:Z** nucleobase pair,[15] which suggested that the presence of the **AEGIS** nucleobases had little impact on duplex structure.

These MD simulations do, however, suggest that bacteria will be able to tolerate the inclusion of **P:Z** nucleobase pairs in their genome given that **PZ**-containing DNA duplexes can adopt both A- and B-form DNA. Thus, the presence of well-defined major and minor grooves will permit interactions with DNA-binding proteins involved in controlling transcriptional replication. In addition, the ability to adopt A-form structures will facilitate DNA replication in engineered DNA polymerases, and the similarity in hydrogen bond interaction energies for G:C and **P:Z** nucleobase pairs makes formation of transcription "bubbles" energetically feasible for **PZ**-containing DNA. Perhaps most importantly, these simulations show that the DNA duplex can be maintained in water even for sequences composed of multiple consecutive **P:Z** nucleobase pairs.

CONCLUSION AND OUTLOOK

This Account has provided an overview of the properties of **AEGIS** nucleobases and **AEGIS** DNA highlighting our most recent crystallographic and computational studies of **PZ**-containing DNA and relevant natural control sequences. Collectively, our studies provide the

first comprehensive analysis of the structural and biophysical properties of artificial DNA containing multiple **P:Z** pairs in a six nucleotide genetic alphabet. **P:Z** pairs exhibit structural features similar to G:C pairs in the same context in both B- and A-DNA crystal structures suggesting that they would be expected to behave similarly in biological reactions. Notable differences include the unusual electrostatic and stacking properties of the **P:Z** nucleobase pairs, which we suggest contribute substantially to the observed dynamical differences between the duplex, 5'-CTTAT**PPPZZZ**ATAAG, and the GC control. Through optimization of DNA polymerase by directed evolution, it is possible to faithfully and efficiently replicate DNA including **P:Z** pairs in PCR reactions.[9] Thus, all of the studies to date support the premise that including **P:Z** pairs expands the genetic alphabet and in so doing the fundamental properties of the DNA.

Biographical Information

Nigel G. J. Richards is Professor of Biological and Organic Chemistry at Cardiff University, UK. He is also a Research Fellow at the Foundation for Applied Molecular Evolution based in Alachua, FL.

Millie M. Georgiadis is Associate Chair and Professor in the Department of Biochemistry and Molecular Biology at Indiana University School of Medicine and has a joint appointment in the Department of Chemistry and Chemical Biology at Indiana University Purdue University at Indianapolis.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the efforts of a number of talented post-doctoral and graduate student researchers, who contributed to both the experimental and computational studies (especially Robert Molt and Isha Singh) outlined herein as well as Dr. Steven Benner (FAME), who created **AEGIS**. This work was supported in part by an NIH grant R01 GM055026 to MMG and funds provided by Cardiff University and Indiana University Purdue University Indianapolis to NGJR.

REFERENCES

(1) Ball, P. Synthetic biology: starting from scratch. *Nature* **2004**, *431*, 624-626.

(2) Benner, S. A. Understanding nucleic acids using synthetic chemistry. *Acc. Chem. Res.* **2004**, *37*, 784-797.

(3) Malyshev, D. A.; Romesberg, F. E. The expanded genetic alphabet. *Angew. Chem. Int. Ed. Engl.* **2015**, *54*, 11930-11944.

(4) Hirao, I.; Harada, Y.; Kimoto, M.; Mitsui, T.; Fujiwara, T.; Yokoyama, S. A two-unnatural-base-pair system toward the expansion of the genetic code. *J. Am. Chem. Soc.* **2004**, *126*, 13298-13305.

(5) Hirao, I.; Kimoto, M.; Yamashige, R. Natural versus artificial creation of base pairs in DNA: origin of nucleobases from the perspectives of unnatural base pair studies. *Acc. Chem. Res.* **2012**, *45*, 2055-2065.

(6) Kool, E. T. Replacing the nucleobases in DNA with designer molecules. *Acc. Chem. Res.* **2002**, *35*, 936-943.

(7) Geyer, C. R.; Battersby, T. R.; Benner, S. A. Nucleobase Pairing in Expanded Watson-Crick-like Genetic Information Systems. *Structure* **2003**, *11*, 1485-1498.

(8) Yang, Z.; Chen, F.; Chamberlin, S. G.; Benner, S. A. Expanded genetic alphabets in the polymerase chain reaction. *Angew. Chem. Int. Ed. Engl.* **2010**, *49*, 177-180.

(9) Laos, R.; Shaw, R.; Leal, N. A.; Gaucher, E.; Benner, S. Directed evolution of polymerases to accept nucleotides with nonstandard hydrogen bond patterns. *Biochemistry* **2013**, *52*, 5288-5294.

(10) Wang, X.; Hoshika, S.; Peterson, R. J.; Kim, M. J.; Benner, S. A.; Kahn, J. D. Biophysics of Artificially Expanded Genetic Information Systems. Thermodynamics of DNA Duplexes Containing Matches and Mismatches Involving 2-Amino-3-nitropyridin-6-one (Z) and Imidazo[1,2-a]-1,3,5-triazin-4(8H)one (P). *ACS Synth Biol* **2017**.

(11) Sismour, A. M.; Lutz, S.; Park, J. H.; Lutz, M. J.; Boyer, P. L.; Hughes, S. H.; Benner, S. A. PCR amplification of DNA containing non-standard base pairs by variants of reverse transcriptase from Human Immunodeficiency Virus-1. *Nucleic Acids Res.* **2004**, *32*, 728-735.

(12) Voegel, J. J.; von Krosigk, U.; Benner, S. A. Synthesis and tautomeric equilibrium of 6-amino-5-benzyl-3-methylpyrazin-2-one. An acceptor-donor-donor nucleoside base analog. *The Journal of Organic Chemistry* **1993**, *58*, 7542-7547.

(13) Sepiol, J.; Kazimierczuk, Z.; Shugar, D. Tautomerism of isoguanosine and solvent-induced keto-enol equilibrium. *Z. Naturforsch* **1976**, *31c*, 361-370.

(14) Chawla, M.; Credendino, R.; Chermak, E.; Oliva, R.; Cavallo, L. Theoretical Characterization of the H-Bonding and Stacking Potential of Two Nonstandard Nucleobases Expanding the Genetic Alphabet. *J. Phys. Chem. B* **2016**, *120*, 2216-2224.

(15) Wang, W.; Sheng, X.; Zhang, S.; Huang, F.; Sun, C.; Liu, J.; Chen, D. Theoretical characterization of the conformational features of unnatural oligonucleotides containing a six nucleotide genetic alphabet. *Phys. Chem. Chem. Phys.* **2016**, *18*, 28492-28501.

(16) Malyshev, D. A.; Dhami, K.; Lavergne, T.; Chen, T.; Dai, N.; Foster, J. M.; Correa, I. R., Jr.; Romesberg, F. E. A semi-synthetic organism with an expanded genetic alphabet. *Nature* **2014**, *509*, 385-388.

(17) Zhang, Y.; Lamb, B. M.; Feldman, A. W.; Zhou, A. X.; Lavergne, T.; Li, L.; Romesberg, F. E. A semisynthetic organism engineered for the stable expansion of the genetic alphabet. *Proc Natl Acad Sci U S A* **2017**, *114*, 1317-1322.

(18) Gould, I. R.; Kollman, P. A. Theoretical Investigation of the Hydrogen Bond Strengths in Guanine-Cytosine and Adenine-Thymine Base Pairs. *J. Am. Chem. Soc.* **1994**, *116*, 2493-2499.

(19) Šponer, J.; Jurečka, P.; Hobza, P. Accurate Interaction Energies of Hydrogen-Bonded Nucleic Acid Base Pairs. *J. Am. Chem. Soc.* **2004**, *126*, 10142-10151.

(20) Jorgensen, W. L.; Pranata, J. Importance of secondary interactions in triply hydrogen bonded complexes: guanine-cytosine vs uracil-2,6-diaminopyridine. *J. Am. Chem. Soc.* **1990**, *112*, 2008-2010.

(21) Molt, R. W.; Georgiadis, M. M.; Richards, N. G. Consecutive Non-Natural PZ Nucleobase Pairs in DNA Impact Helical Structure as Seen in 50 ms Molecular Dynamics Simulations. *doi.org/10.1093/nar/gkx144* **2017**.

(22) Ashworth, J.; Havranek, J. J.; Duarte, C. M.; Sussman, D.; Monnat, R. J., Jr.; Stoddard, B. L.; Baker, D. Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* **2006**, *441*, 656-659.

(23) Mooers, B. H. Crystallographic studies of DNA and RNA. *Methods* **2009**, *47*, 168-176.

(24) Wing, R.; Drew, H.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K.; Dickerson, R. E. Crystal structure analysis of a complete turn of B-DNA. *Nature* **1980**, *287*, 755-758.

(25) Cote, M. L.; Yohannan, S. J.; Georgiadis, M. M. Use of an N-terminal fragment from moloney murine leukemia virus reverse transcriptase to facilitate crystallization and analysis of a pseudo-16-mer DNA molecule containing G-A mispairs. *Acta Crystallogr D Biol Crystallogr* **2000**, *56*, 1120-1131.

(26) Cote, M. L.; Pflomm, M.; Georgiadis, M. M. Staying straight with A-tracts: a DNA analog of the HIV-1 polypurine tract. *J. Mol. Biol.* **2003**, *330*, 57-74.

(27) Montano, S. P.; Cote, M. L.; Roth, M. J.; Georgiadis, M. M. Crystal structures of oligonucleotides including the integrase processing site of the Moloney murine leukemia virus. *Nucleic Acids Res.* **2006**, *34*, 5353-5360.

(28) Goodwin, K. D.; Lewis, M. A.; Tanious, F. A.; Tidwell, R. R.; Wilson, W. D.; Georgiadis, M. M.; Long, E. C. A high-throughput, high-resolution strategy for the study of site-selective DNA binding agents: analysis of a "highly twisted" benzimidazole-diamidine. *J. Am. Chem. Soc.* **2006**, *128*, 7846-7854.

(29) Cote, M. L.; Georgiadis, M. M. Structure of a pseudo-16-mer DNA with stacked guanines and two G-A mispairs complexed with the N-terminal fragment of Moloney murine leukemia virus reverse transcriptase. *Acta Crystallogr D Biol Crystallogr* **2001**, *57*, 1238-1250.

(30) Georgiadis, M. M.; Singh, I.; Kellett, W. F.; Hoshika, S.; Benner, S. A.; Richards, N. G. Structural basis for a six nucleotide genetic alphabet. *J. Am. Chem. Soc.* **2015**, *137*, 6947-6955.

(31) Singh, I.; Jian, Y.; Li, L.; Georgiadis, M. M. The structure of an authentic spore photoproduct lesion in DNA suggests a basis for recognition. *Acta Crystallogr D Biol Crystallogr* **2014**, *70*, 752-759.

(32) Goodwin, K. D.; Lewis, M. A.; Long, E. C.; Georgiadis, M. M. Crystal structure of DNA-bound Co(III) bleomycin B2: Insights on intercalation and minor groove binding. *Proc Natl Acad Sci U S A* **2008**, *105*, 5052-5056.

(33) Glass, L. S.; Nguyen, B.; Goodwin, K. D.; Dardonville, C.; Wilson, W. D.; Long, E. C.; Georgiadis, M. M. Crystal structure of a trypanocidal 4,4'-bis(imidazolinylamino)diphenylamine bound to DNA. *Biochemistry* **2009**, *48*, 5943-5952.

(34) Goodwin, K. D.; Long, E. C.; Georgiadis, M. M. A host-guest approach for determining drug-DNA interactions: an example using netropsin. *Nucleic Acids Res.* **2005**, *33*, 4106-4116.

(35) Burge, S.; Parkinson, G. N.; Hazel, P.; Todd, A. K.; Neidle, S. Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.* **2006**, *34*, 5402-5415.

(36) Zhang, L.; Yang, Z.; Sefah, K.; Bradley, K. M.; Hoshika, S.; Kim, M. J.; Kim, H. J.; Zhu, G.; Jimenez, E.; Cansiz, S.; Teng, I. T.; Champanhac, C.; McLendon, C.; Liu, C.; Zhang, W.; Gerloff, D. L.; Huang, Z.; Tan, W.; Benner, S. A. Evolution of functional six-nucleotide DNA. *J. Am. Chem. Soc.* **2015**, *137*, 6734-6737.

(37) Lu, X. J.; Olson, W. K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **2003**, *31*, 5108-5121.

(38) Jurečka, P.; Hobza, P. True Stabilization Energies for the Optimal Planar Hydrogen-Bonded and Stacked Structures of Guanine···Cytosine, Adenine···Thymine, and Their 9- and 1-Methyl Derivatives: Complete Basis Set Calculations at the MP2 and CCSD(T) Levels and Comparison with Experiment. *J. Am. Chem. Soc.* **2003**, *125*, 15608-15613.

(39) Wang, A. H.; Fujii, S.; van Boom, J. H.; Rich, A. Molecular structure of the octamer d(G-G-C-C-G-G-C-C): modified A-DNA. *Proc Natl Acad Sci U S A* **1982**, *79*, 3968-3972.

(40) Rohs, R.; Jin, X.; West, S. M.; Joshi, R.; Honig, B.; Mann, R. S. Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem* **2010**, *79*, 233-269.

(41) Cheatham, T. E., 3rd; Case, D. A. Twenty-five years of nucleic acid simulations. *Biopolymers* **2013**, *99*, 969-977.

(42) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E., 3rd; Laughton, C. A.; Orozco, M. Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.* **2007**, *92*, 3817-3829.

(43) Zgarbova, M.; Otyepka, M.; Sponer, J.; Mladek, A.; Banas, P.; Cheatham, T. E., 3rd; Jurecka, P. Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J Chem Theory Comput* **2011**, *7*, 2886-2902.

(44) MacKerell, A. D., Jr.; Banavali, N.; Foloppe, N. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* **2000**, *56*, 257-265.
(45) Joung, I. S.; Cheatham, T. E., 3rd. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **2008**, *112*, 9020-9041.
(46) Galindo-Murillo, R.; Robertson, J. C.; Zgarbova, M.; Sponer, J.; Otyepka, M.; Jurecka, P.; Cheatham, T. E., 3rd. Assessing the Current State of Amber Force Field Modifications for DNA. *J Chem Theory Comput* **2016**, *12*, 4114-4127.
(47) Pasi, M.; Maddocks, J. H.; Beveridge, D.; Bishop, T. C.; Case, D. A.; Cheatham, T., 3rd; Dans, P. D.; Jayaram, B.; Lankas, F.; Laughton, C.; Mitchell, J.; Osman, R.; Orozco, M.; Perez, A.; Petkeviciute, D.; Spackova, N.; Sponer, J.; Zakrzewska, K.; Lavery, R. muABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.* **2014**, *42*, 12272-12283.

Figure Legends

Figure 1. Chemical structures of **P**:**Z**, **X**:**K**, **G**:**C** and **A**:**T** nucleobase pairs. R is 2'-deoxyribose in duplex DNA.

Figure 2. Dipole moments (arrows) and electrostatic potential (rendered on the VDW surface) for **P:Z** (left) and **G:C** (right) nucleobase pairs. Electrostatic energies range from approximately -40 kcal/mol (red) to ~ +40 kcal/mol (blue). Taken from Ref. 19 and used with permission.

Figure 3. The host-guest system is shown with host protein molecules (N-terminal fragment of MMLV RT) as ribbon renderings in green and blue and the guest as a cartoon rendering with the two strands of the self-complementary sequence 5'-CTTAT**PP**TA**ZZ**ATAAG in yellow and magenta, **P:Z** pairs in green. The box indicates the contents of the asymmetric unit of the host-guest crystal, one protein molecule and half of the 16-mer DNA duplex. Arrows from the duplex DNA indicate the positions of the **P:Z** pairs and end in a stick-rendering of the **P:Z** pair, top of figure. An arrow from the protein-DNA interface in the complex points to a close-up rendering of this interaction mediated by binding of R116, positioned through hydrogen bonding interactions with D114, to O2 of the second base from the end in the minor groove, shown at the bottom of figure.

Figure 4. P:Z in B-form DNA is sheared by -0.87 Å and G:C by -1.40 Å while the equivalent A:T pair does not exhibit significant shearing (-0.22 Å) as shown on the right-hand side of the figure with stick renderings and filled rings for each of the nucleobase pairs. Shearing defines displacement along the hydrogen-bonding edge of one base with respect to the other. Although sheared, both **P:Z** and G:C retain standard hydrogen bonding distances. On the left are shown the same base pairs in an edge-on view. In this view, it is evident that buckle angle of -11.91° for **P:Z** is much larger than that in G:C or A:T, -5.6°and -1.4°, respectively. The buckle angle defines the degree of non-planarity across the base pair.

Figure 5. Van der Waal renderings of the major and minor groove faces of the **P:Z**, G:C, and A:T base pairs are shown, with N atoms in blue, O in red, P, orange, and C in either yellow for **P:Z**, green for G:C or pink for A:T.

Figure 6. Crystal structure of **P:Z** pairs in A-DNA. A) Stick rendering of the crystal structure of 3/6 ZP with N blue, O, red, P, orange, and C in green for Z, yellow for P, and standard pairs in orange. B) End view of the same structure.

Figure 7. Individual **P:Z** pair from the 3/6ZP A-DNA structure is shown as a stick rendering along with a G:C pair from a G-rich structure in A-form (PDB identifier: 4OKL). These pairs in contrast to those in B-form (depicted in Fig. 4) are not sheared or buckled as shown on the right and left panels, respectively.

Figure 8. Stacking interactions for two **Z:P** pairs are shown in a top down view for the “slide” (a) and (b) and “shift” conformers (c) and (d). Views (a) and (c) are derived from 36ZP A-form and 2P B-form crystal structures, respectively, while (b) and (d) are calculated for isolated

stacked **Z:P** pairs. N5 atoms of the **Z** nitro group are highlighted by encircling in cyan, bonds for the bottom **Z** ring are highlighted in green.

Figure 9. Representative PZ and GC structures from MD simluations. Stick models are shown for the PZ oligonucleotide in an extended conformation (far left), PZ in an A-like conformation (center), and GC in a B-like conformation. Views are shown parallel (upper panels) and perpendicular (lower panels) to the helical axes. In each rendering, N is blue, O red, P orange, and C light gray. Taken from Ref. 19 and used with permission.

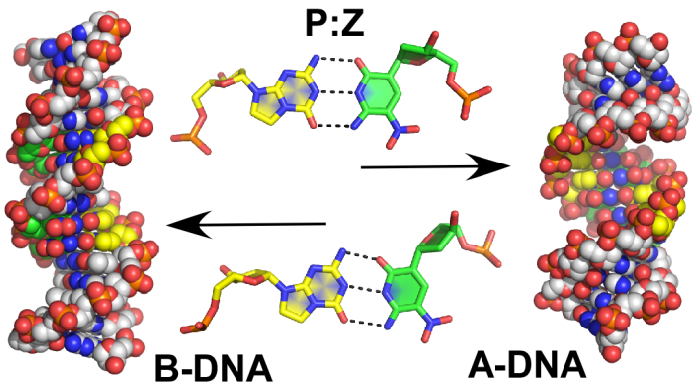


Figure for Abstract.

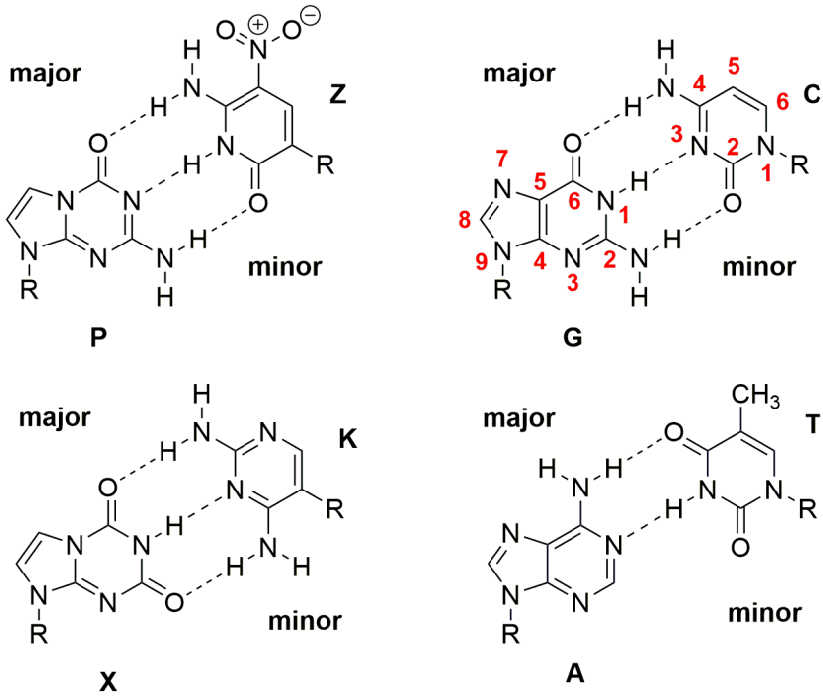
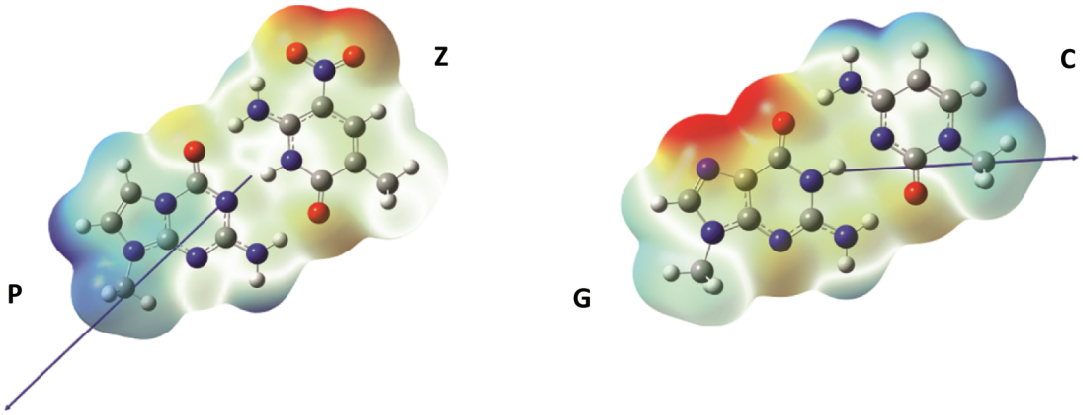
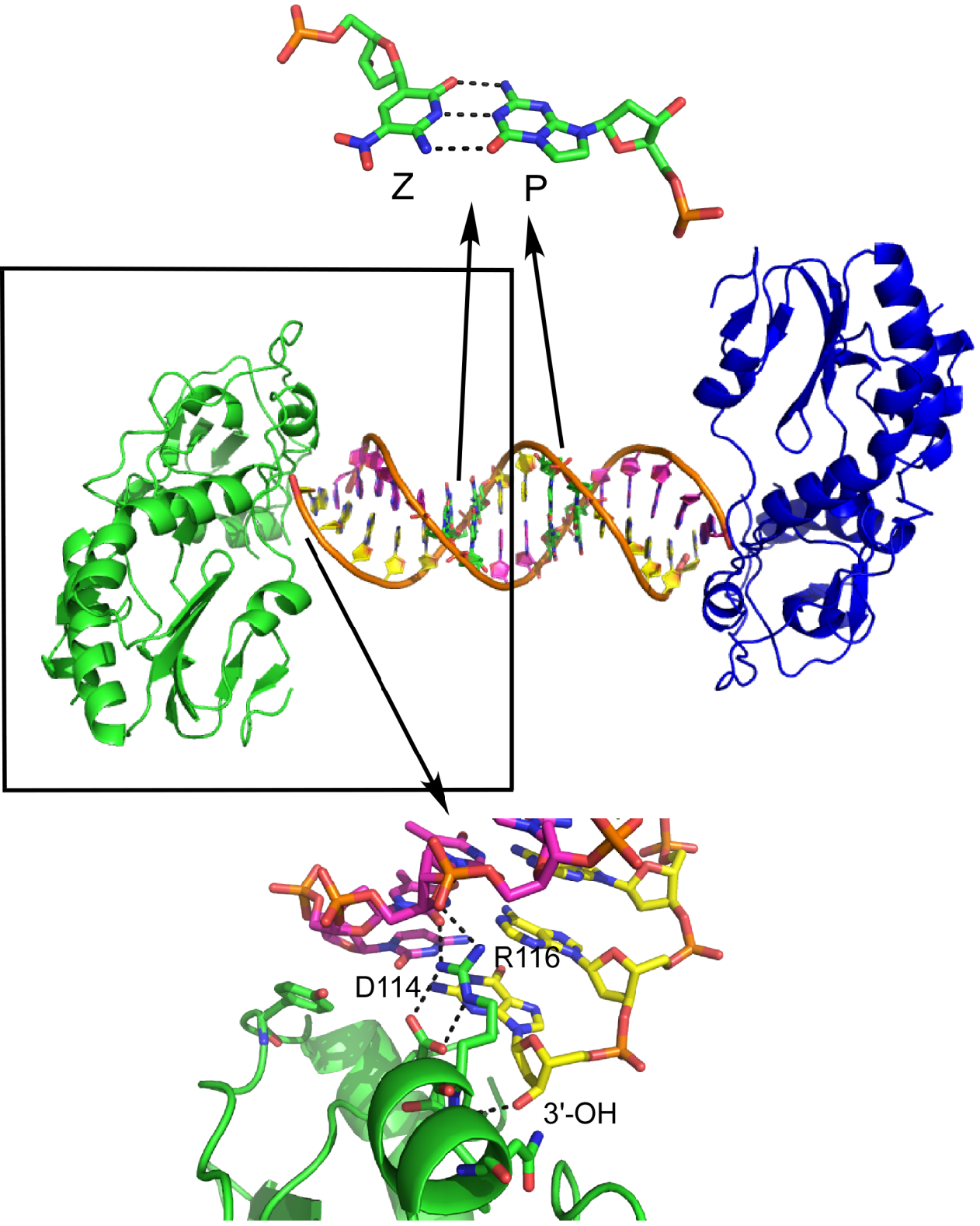


Fig. 1


Z
P
C
G


Z
P
R116
D114
3'-OH

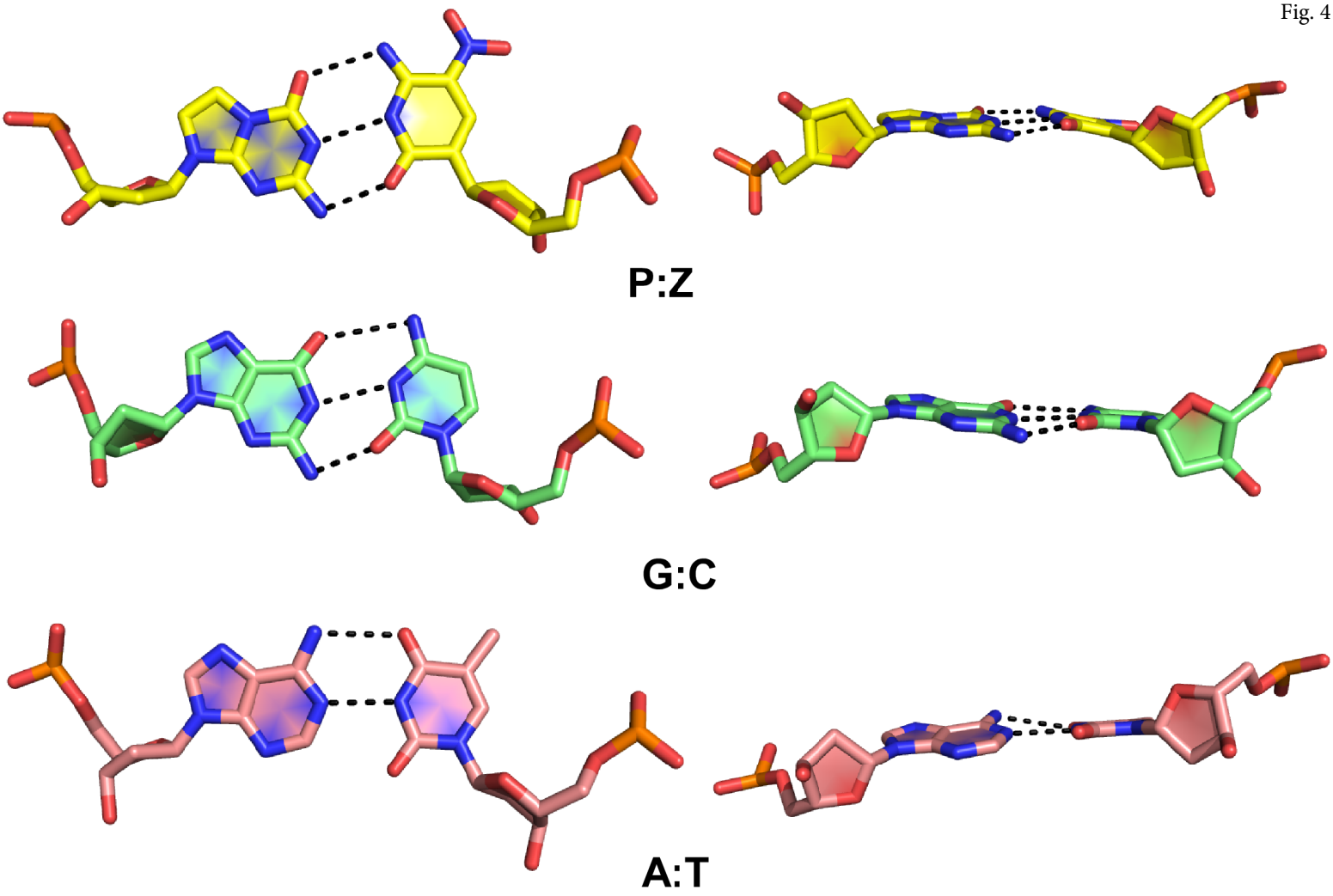
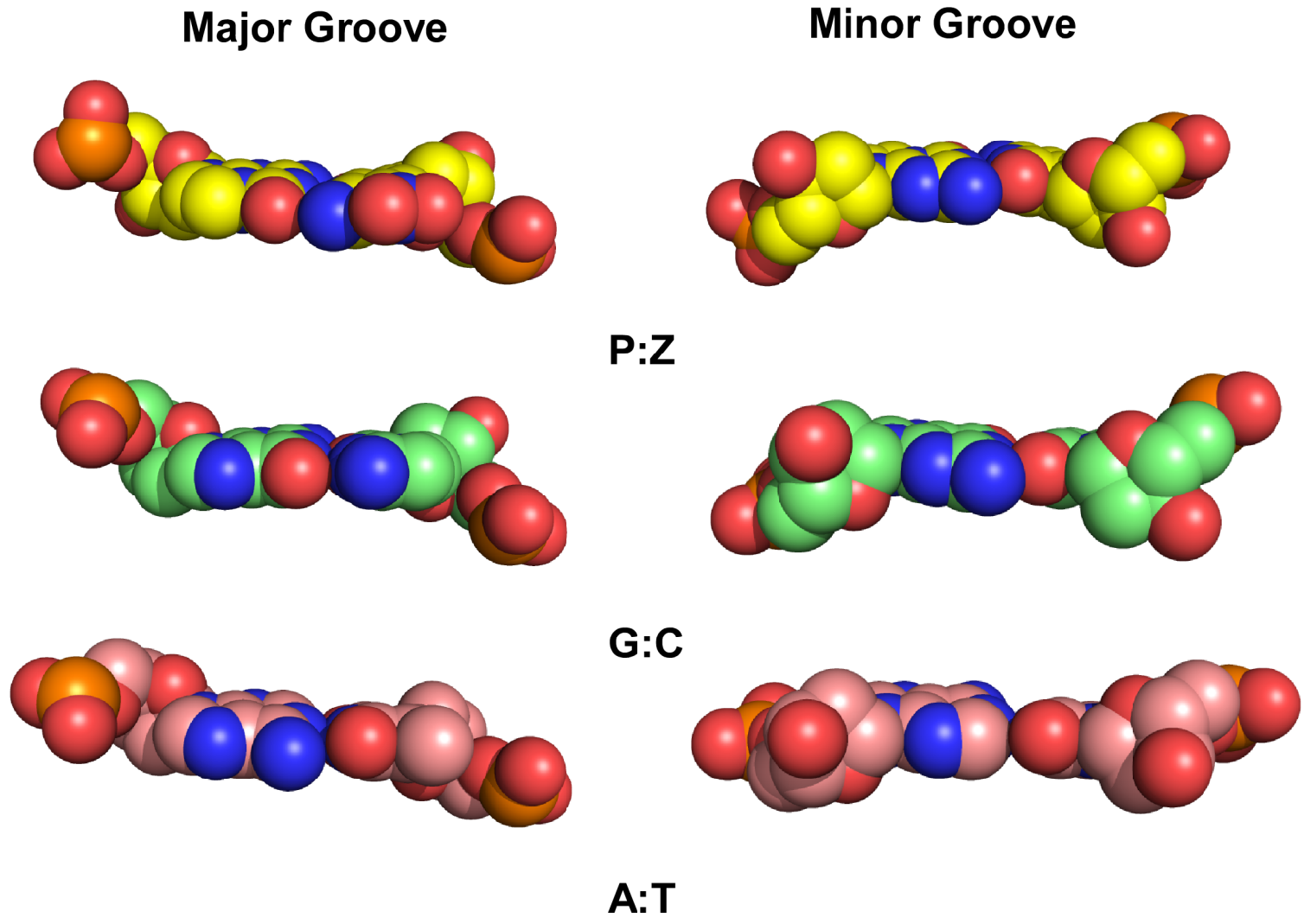

P:Z
G:C
A:T

Fig. 5



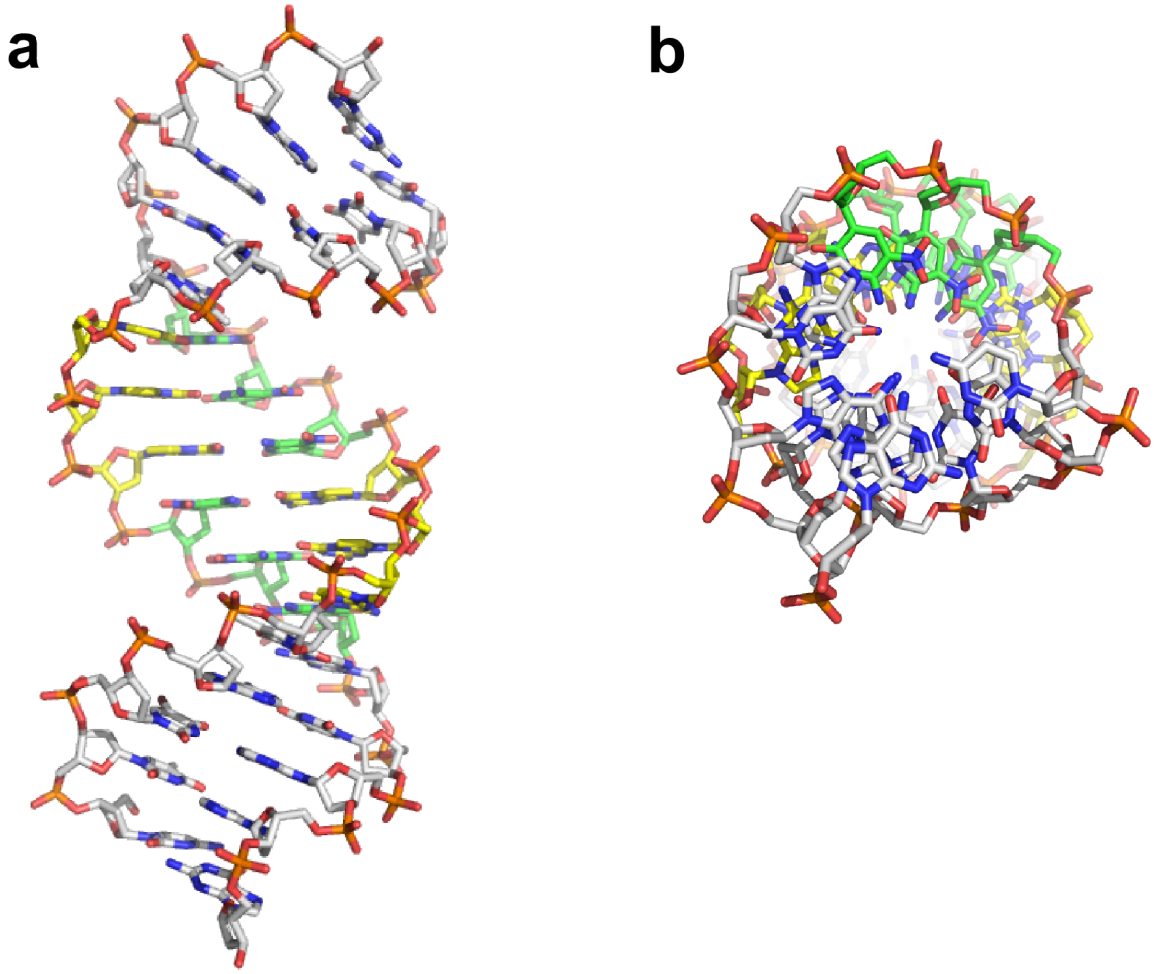
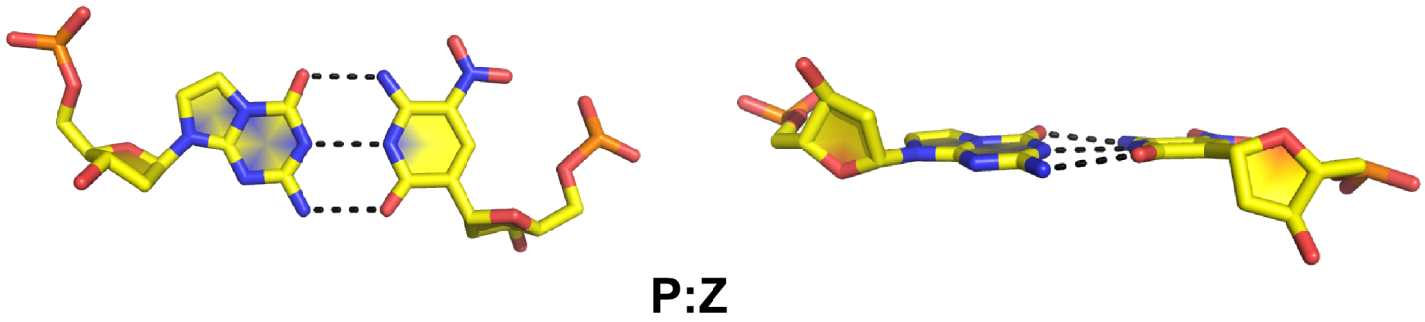
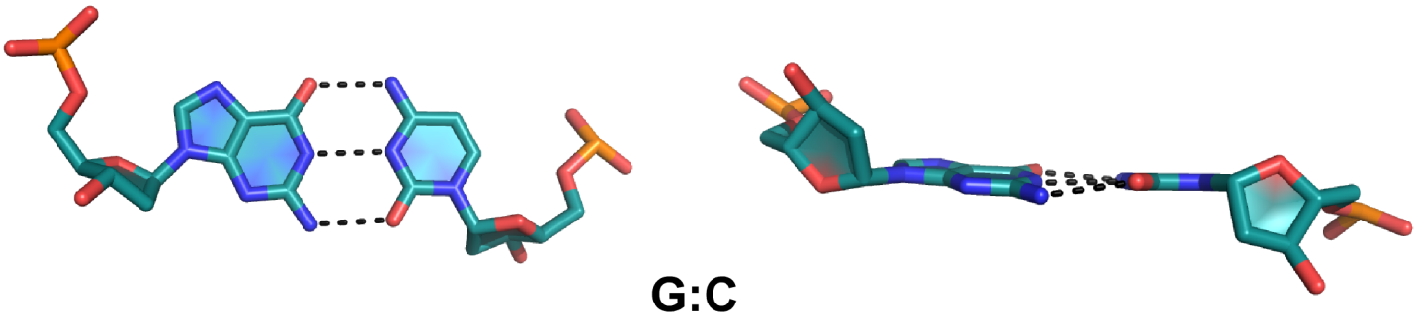


Fig. 6

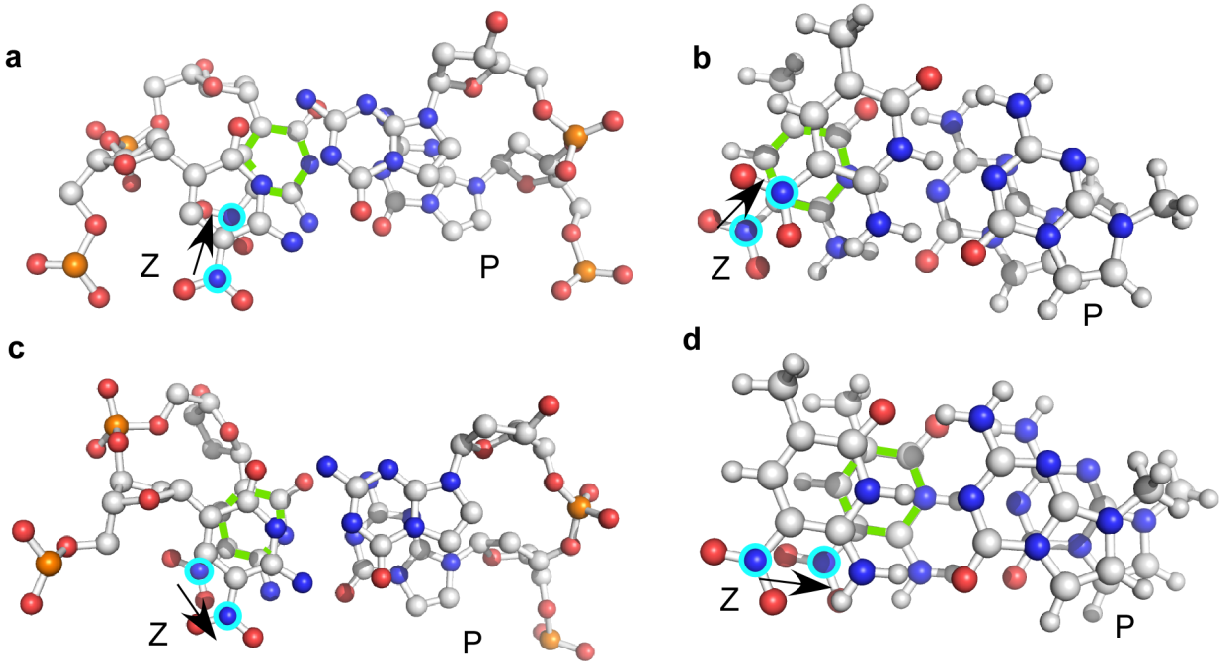


P:Z



G:C

Fig. 8



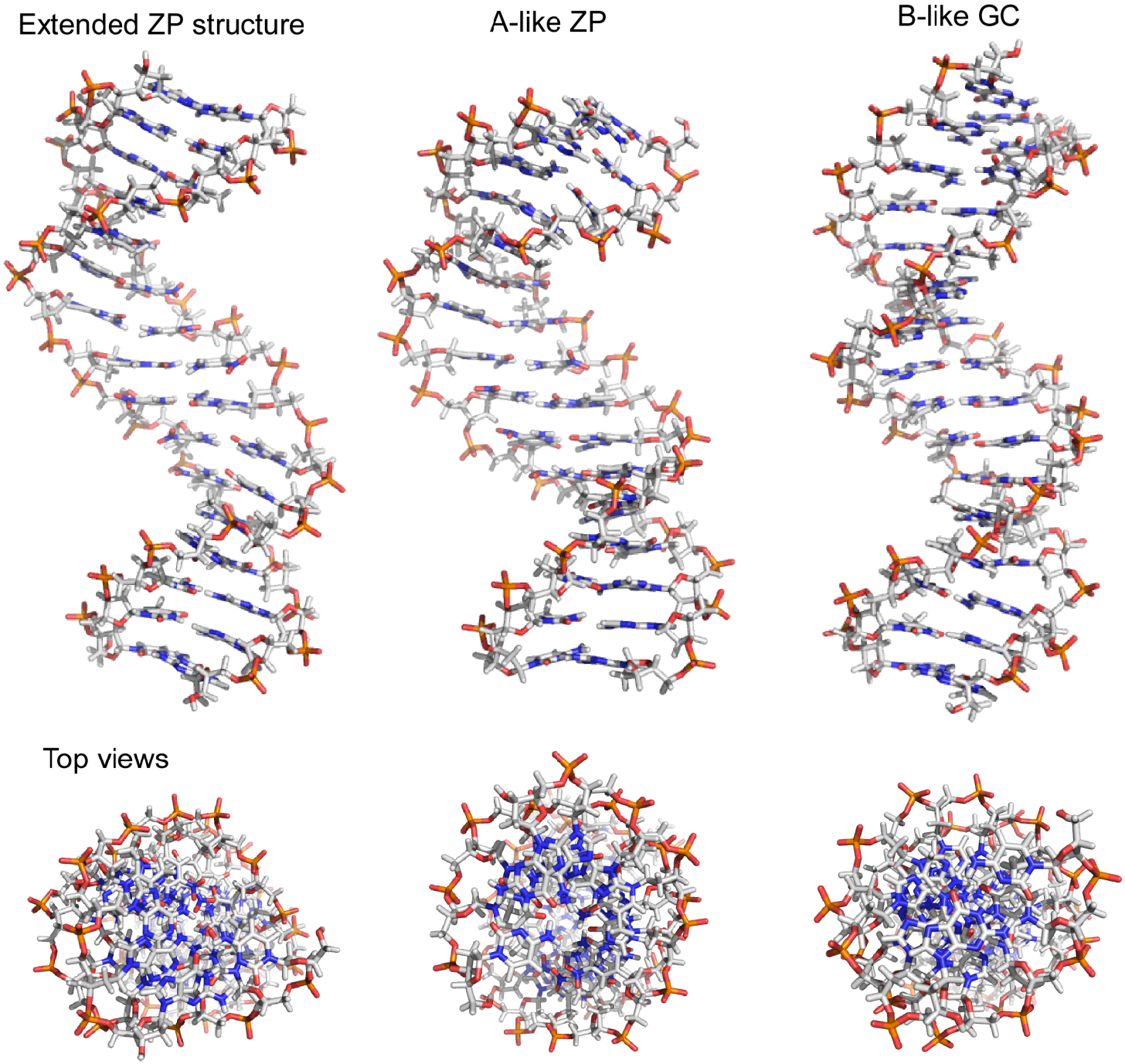


Fig. 9